DERIVING CONFIDENCE SETS FOR EFFECT SIZES USING SIMULTANEOUS CONFIDENCE INTERVALS

By

Kenneth Liao

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

May 10, 2024

Nashville, Tennessee

Approved:

Simon Vandekar, Ph.D.

Jinyuan Liu, Ph.D.

ACKNOWLEDGMENTS

I would like to first thank Dr. Simon Vandekar for introducing me to the field of neuroimaging and explaining the foundations of neuroimaging inference to me. I am grateful for the opportunity to work on a project as interesting as this, and appreciate his flexibility to take time to meet with me every week. His instruction and supplying of neuroimaging literature was integral in my learning of how neuroimaging works. Thank you for your help on this thesis during the last few weeks before the deadline. I would also like to thank Dr. Jinyuan Liu for being the second reader of my thesis, and for providing great insights and questions during my presentation. Thank you for taking the time while you were out of the office to be a member of my committee.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

The structure of the brain plays an integral role in determining the behavioral phenotype of a person. For instance, variations in brain structure like different amounts of gray matter volume can influence cognitive functions and social behaviors. The association between the brain and these behaviors is of interest in neurocognitive and neuropsychiatric research. Brain-wide association studies (BWAS) use noninvasive magnetic resonance imaging (MRI) to identify these associations of brain structure and function with psychometric measurements Marek et al. (2022). After characterizing how differences in brain structure and function are associated with inter-individual differences in cognition or psychiatric symptoms, we can identify how neurological differences may serve as markers for diagnosis that precede formation of persistent and severe symptoms of a disorder. These associations can inform us of illness trajectories, neural mechanisms, and differences between individuals. For instance, the heterogeneity in autism spectrum disorder (ASD) presents a considerable challenge to diagnosis and precision treatment. Neuroimaging analysis can address the issue of heterogeneity by showing how different ASD clinical subtypes such as Asperger's, PDD-NOS (pervasive developmental disorder-not otherwise specified) and Autistic are linked to unique brain systems and subdomain symptoms Qi et al. (2020).

Conducting group-level neuroimaging analysis can help us to identify these brain-behavior associations through hypothesis testing. Neuroimaging inference focuses almost exclusively on hypothesis testing, to detect significantly active regions of the brain Friston et al. (1994) Friston et al. (1996). The most widely used methods include inference on individual voxels of an image, and on clusters of voxels in an image. Voxel-wise inference requires conducting hypothesis testing for each voxel in an image, which can amount up to 100,000s of data points, combined with high-dimensional behavioral data. Cluster extent inference can address this problem by identifying statistically significant clusters of voxels. These two methods are discussed in more detail in section 1.2. The reliance of these methods in reporting p-values for statistical inference is an issue due to its reliance on sample size, as even minuscule effects can be determined as significant with a large enough sample size Wasserstein and Lazar (2016). Methods that generate inferences based on effect sizes have been shown to address these limitations from p-values.

This research aims to create a general, established method for using confidence sets to conduct effect size-based inference.

1.1 Neuroimaging group-level analysis notation

Let $Y_i(v)$ denote the outcome vector image for each subject $i = 1, 2, \ldots, n$, which measures brain activation. All images are indexed by the voxel location, $v \in \mathbb{B} \subset \mathbb{R}^3$, where $\mathbb{B}$ denotes the bounded space of the brain. We fit the model

$$\begin{aligned} Y_i(v) &= X_{i0}\alpha(v) + X_{i1}\beta(v) + E_i(v) \\ &= X_i\zeta(v) + E_i(v) \end{aligned} \tag{1.1}$$

where $X_{i0} \in \mathbb{R}^{m_0}$ is a row vector of nuisance covariates including the intercept, $X_{i1} \in \mathbb{R}^{m_1}$ is a row vector of diagnostic variables of interest, $m = m_0 + m_1$, and $X_i = [X_{i0}, X_{i1}] \in \mathbb{R}^m$; $\alpha(v)$ and $\beta(v)$ are parameter image vectors that take values in $\mathbb{R}^{m_0}$ and $\mathbb{R}^{m_1}$, respectively, and $\zeta(v) = [\alpha(v)^T, \beta(v)^T]^T \in \mathbb{R}^m$; $E_i(v)$ is an error term vector with mean zero and spatial covariance matrix $\Sigma_i(v,w) = Cov\{E_i(v), E_i(w)\}$ for any two voxels v and w. This covariance describes the dependence between repeated imaging measurements and spatially within the image. The multilevel and spatial aspects of the data make analyses susceptible to unknown multivariate heteroscedasticity: $\Sigma_i(v,w) \neq \Sigma_j(v,w)$ for $i \neq j$.

1.2 Hypothesis testing in neuroimaging

Cluster extent inference (CEI), also termed spatial extent inference (SEI), is one of the most common inference procedures in neuroimaging used to identify how functional and anatomical networks are associated with a covariate such as diagnosis or acute symptomology Woo et al. (2014). After fitting the linear model 1.1, we can use CEI to conduct hypothesis testing on $\beta(v)$, the quantification of group differences in activation related to cognitive demand at location v. We first specify a null hypothesis for the entire image,

$$H_0(v) : \beta(v) = 0, \forall v \in \mathbb{B}. \tag{1.2}$$

Let $Z(v)$ denote the Chi-squared statistical image (Figure 1.1a) for the test of (1.2) and associated p-value image $p(v)$. The statistical image $Z(v)$ is then thresholded (Figure 1.1b) at a cluster forming threshold (CFT), p_0, which denotes an uncorrected p-value threshold. For example, if $p_0 = 0.01$, then the Chi-squared statistical image $Z(v)$ is thresholded at $\Phi_1^{-1}(1-0.01) = 6.63$, where Φ_1^{-1} denotes the quantile function of a Chi-squared distribution, so that locations v where the observed image $Z(v) > 6.63$ have $p(v) < 0.01$. This thresholding thus identifies spatially contiguous suprathreshold clusters, which are binarized (Figure 1.1c) and the spatial extent (number of voxels within each cluster) can be obtained (Figure 1.1d). Finally, CEI will compute a p-value for each cluster based on the image null hypothesis 1.2.

However, for any sample size, CEI only indicates regions where a null hypothesis can be rejected, without providing any notion of spatial uncertainty about the activation. For instance, we can only conclude that brain

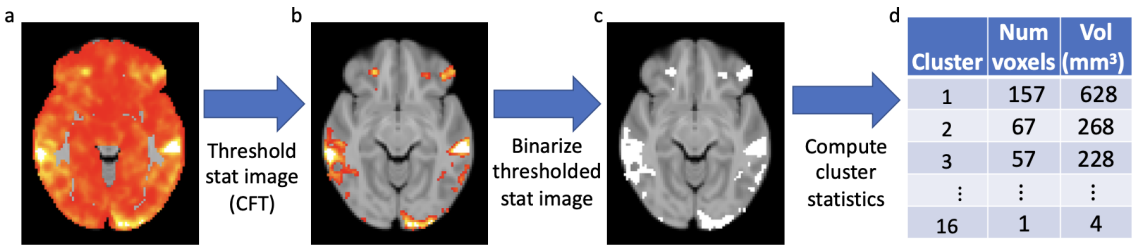

Figure 1.1: This is a 2-dimensional illustration of how CEI is performed. (b) illustrates the statistical image of the brain (a) after being thresholded. (c) shows the statistical image after the clusters from (b) are binarized, and tabulated in (d), ordered by decreasing number of voxels.

activation has occured somewhere inside a given cluster because the significance of specific voxels cannot be determined. Additionally, if a cluster covered multiple anatomical regions of the brain, we can not pinpoint the precise source of activation. The larger a cluster gets, the spatial specificity of the inference will diminish. CEI also lacks information on spatial variability. If we were to repeat a study multiple times with different groups of people, there will be variation in the sizes and shapes of the final activation clusters, but CEI can not convey this. Bowring et al. (2019)

The focus on detection of non-zero signal or signal change is a problem that is exacerbated for large scale studies, where the ”null hypothesis fallacy” can cause even trivial effects to be determined as significant. This is because statistical models conventionally assume mean-zero noise, but in reality, there will always be some non-zero signal everywhere because all sources of noise can never completely cancel. Bowring et al. (2021) Thus, with increasing sample size, the smallest of effects will eventually become statistically significant.

1.2.1 Voxel-wise inference

In contrast to conducting inference on regions of the brain, we can focus inference to the voxels of an image of the brain. To account for the large amount of tests needed to be done for each voxel, we can control the familywise error rate (FWER) or false discovery rate (FDR), through selecting some effect size threshold so that the probability of at least one null hypothesis being falsely rejected is less than or equal to some specified α Vandekar et al. (2018).

Parametric methods for imaging analyses have been shown to fail, yielding conservative FWERs for voxelwise inference and invalid FWERs for CEI. To account for this inflation of false-positive rates, there have been transitions towards use of the nonparametric permutation test and bootstrapping to perform inference instead Eklund et al. (2016).

1.3 Confidence set approaches

To address the issue of spatial uncertainty from CEI, spatial confidence sets (CSs) can be developed on clusters found in thresholded Cohen's d effect size images Bowring et al. (2021), illustrated in Figure 1.2. This is an improvement from typical hypothesis testing, where the only the null, i.e. a raw effect size of zero, can be rejected. From a non-zero threshold, the constructed upper (target) and lower (null) CS can give statements on where effect sizes exceed or fall short of the threshold, respectively. Bowring et al. (2019) This method can go beyond statistical hypothesis testing as we can make inference not only on regions of the brain that have responded to a task, but also on regions that did not respond to a task. Note that test statistics are not suitable to create CSs as they do not estimate population quantities.

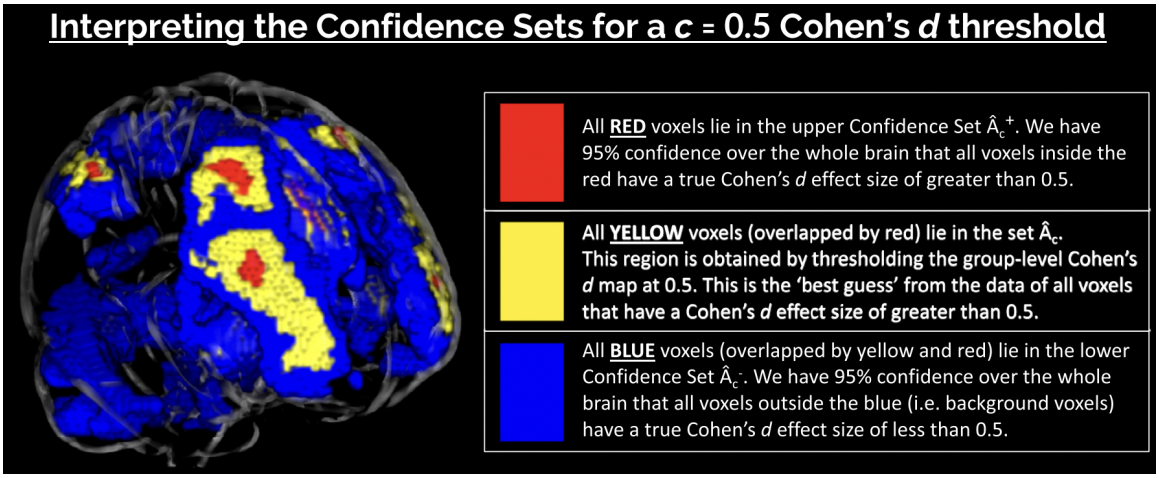

Figure 1.2: This is a 2-dimensional illustration to help visualize how confidence sets are formed for an image of the brain. Here, the red voxels indicate our target set, where we have 95% confidence that all voxels in this area have a true effect size greater than a prespecified threshold. This is a subset of the yellow voxels, which is the point estimate set. Finally, the red and yellow voxels are a subset of the blue voxels (outer set); we have 95% confidence that all voxels outside of the blue have a true effect size less than the threshold. Adapted from Bowring et al. (2021).

The issue with using effect sizes such as Cohen's d is that they parameter specific, model specific, and rely on the model being completely accurate. Cohen's $d = \frac{\mu_0 - \mu_1}{\sigma}$, where μ_0 and μ_1 denote the population means for each group and σ denotes their shared standard deviation. This shared standard deviation σ makes the assumption that the variance is equal across groups and does not explicitly allow for covariates.

Additionally, these CS methods have been heavily focused on functional and continuous data, such as imaging data. Thus, there is a need to make these CS methods generally defined for all types of statistics.

Methods have been developed to derive confidence sets from simultaneous confidence intervals (SCIs) Ren et al. (2023). SCIs differ from confidence intervals (CIs) in that they are able to capture every parameter at a prespecified confidence level. For instance, if we have ten parameters of interest with a prespecified confidence level of 0.95, we can have 95% confidence that the SCIs capture all ten parameters at the same

time, while traditional CIs are not guaranteed to capture all ten simultaneously. We can expect the width of SCIs to be wider than traditional confidence intervals because of this additional requirement to capture every single parameter. Thus, we aim to develop a general methods to obtain CSs for effect sizes through the use of SCIs. A simple 1-dimensional illustration of this procedure is shown in Figure 1.3.

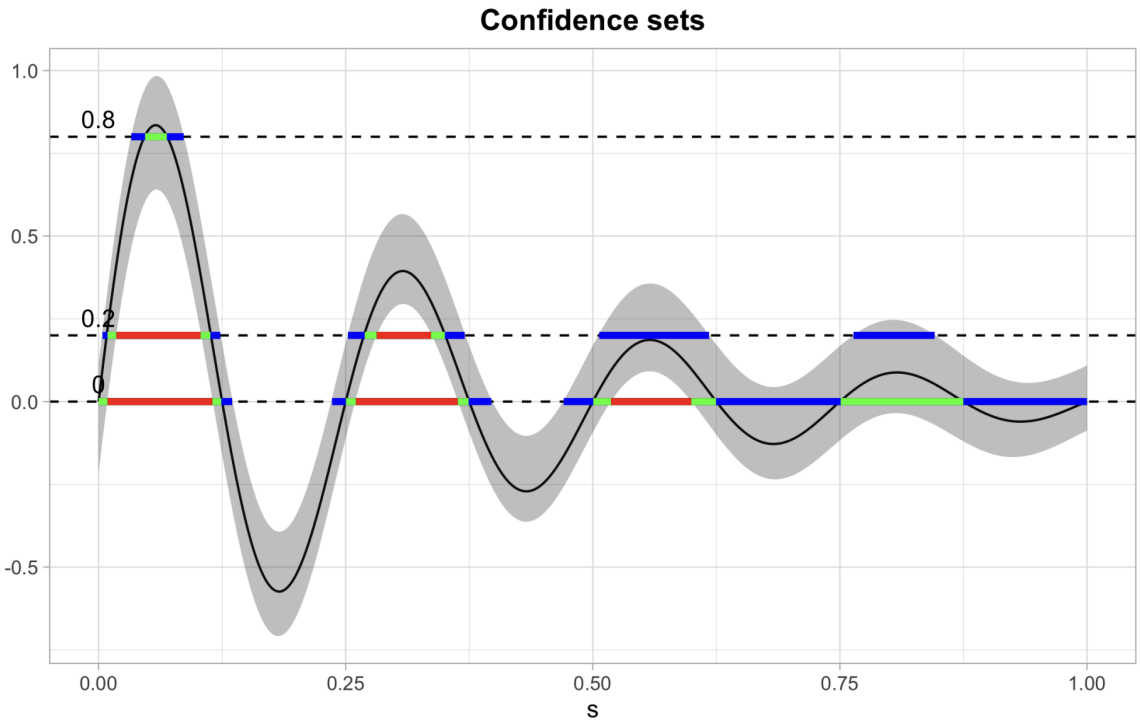

Figure 1.3: This graph displays a 1-dimensional example of how confidence sets can be derived from SCIs, where $\alpha = 0.05$. The location in the 1-D image is shown on the x-axis, and the value of the effect size is shown on the y-axis, thresholded at three different values 0, 0.2, and 0.8. The grey area is the 95% simultaneous confidence band (SCB) formed by the upper and lower SCI, and the solid black line is the true effect size. Regions of the image outlined by the red horizontal lines are contained in the target set, where we have 95% confidence that the true effect size is greater than the given threshold. This is a subset of the green lines, which are the point estimate set. The red and green are a subset of the blue lines, indicating the outer set, where we have 95% confidence that regions outside the blue have a true effect size below the threshold. Adapted from Ren et al. (2023).

1.4 Robust effect size index

Although the use of confidence sets has been developed for effect sizes, these methods are not able to be generalized to all statistics. A recently developed framework for effect size-based inference uses a robust effect size index (RESI), which is generally defined across many different types of models. It is termed as "robust" because its estimator is consistent under model misspecification when estimated with a robust test statistic Vandekar et al. (2020).

The RESI is generally defined across many different types of models because an estimator can be com-

puted from any test statistic, and can be used in a framework for effect-size based inference Jones et al. (2023). We use this RESI framework to develop a general approach to effect size-based inference using confidence sets.

1.4.1 RESI definition

Let $W = \{W_1, ..., W_n\}$ denote the full dataset of independent observations, and let $\theta = (\alpha, \beta)$ be a vector of parameters as defined in 1.1. The RESI is defined from the test statistic for the null hypothesis $H_0 : \beta = \beta_0$, where β_0 is a vector-valued reference value that is usually zero Vandekar and Stephens (2021). The typical Wald-style statistic for the test of the null hypothesis follows a Chi-squared distribution with m_1 degrees of freedom,

$$T^2_{n,m_1} = n(\hat{\beta} - \beta_0)^T \Sigma_\beta^{-1}(\theta)(\hat{\beta} - \beta_0) \sim \chi^2_{m_1}\{n(\beta - \beta_0)^T \Sigma_\beta^{-1}(\theta)(\beta - \beta_0)\}, \tag{1.3}$$

where $\Sigma_\beta(\theta)$ is the asymptotic covariance matrix of $\sqrt{n}(\hat{\beta} - \beta)$ and $\hat{\beta}$ is the estimated value of β Kang et al. (2023). The RESI is defined as the square root of the component of the noncentrality parameter that does not depend on the sample size,

$$S_\beta = \sqrt{(\beta - \beta_0)^T \Sigma_\beta^{-1}(\beta - \beta_0)}. \tag{1.4}$$

CHAPTER 2

Methods

2.1 Effect Size Estimation

To estimate the true effect size

$$S(v) = \sqrt{\frac{\beta(v)^2}{\text{Var}(\sqrt{n}\hat{\beta}(v))}}, \tag{2.1}$$

where

$$\text{Var}(\sqrt{n}\hat{\beta}(v)) = \frac{1}{p(1-p)}, \tag{2.2}$$

we use the estimator

$$\tilde{S}(v) = \sqrt{\tilde{S}^2(v)} = \sqrt{T^2(v)/n}, \tag{2.3}$$

where $T^2(v)$ is the Chi-squared statistical image. This is different than our usual estimator for $S(v)$,

$$\hat{S}(v) = \sqrt{\max\{(T^2(v)-d)/n, 0\}}, \tag{2.4}$$

where d is the degrees of freedom of T. We use $\tilde{S}^2(v)$ instead of $\hat{S}(v)$ because it is linear in the Chi-squared or F statistic, so it is easier to calculate its variance.

2.1.1 Algorithm

To obtain the SCIs for the effect sizes of each voxel, we implement a non-parametric bootstrap.

The bootstrap procedure, modified from Ren et al. (2023), is as follows:

1. For each bootstrap:
   (a) Obtain $Y(v)_b$ from the non-parametric bootstrap, and then estimate the effect size, $\tilde{S}(v)_b$. We also obtain $\hat{\sigma}(v)$, which is defined below in step 4.
   (b) Normalize the test statistic by subtracting by the RESI estimate computed in the observed data, $\tilde{S}(v)$, and dividing by the standard deviation of $\tilde{S}(v)_b$, $\hat{\sigma}(v)_b$.
   (c) Take the maximum and minimum element of $\frac{\tilde{S}(v)_b - \tilde{S}(v)}{\hat{\sigma}(v)_b}$ and append them to vectors $\mathbf{w}^{max}$ and $\mathbf{w}^{min}$, respectively.
2. Denote the upper quantile, cu, as the $1-\frac{\alpha}{2}$ quantile of $\mathbf{w}^{max}$.

3. Denote the lower quantile, cl, as the $\frac{\alpha}{2-\alpha}$ quantile of the subset of $\mathbf{w}^{min}$, at the indices where $\mathbf{w}^{max} \leq cu$. Gao et al. (2021)

4. $$\text{SCI} = (\tilde{S}(v) - cu * \hat{\sigma}(v), \tilde{S}(v) - cl * \hat{\sigma}(v)), \tag{2.5}$$

   where $\hat{\sigma}(v))$ is the standard deviation of the RESI estimate computed in the observed data. This formula of the SCI was derived from Hall (2013).

2.2 Normalization Methods

We consider two different methods of normalizing the test statistic. Assuming the test statistic is F-distributed, we approximate the standard deviation of $\tilde{S}(v)$ using the variance of $\sqrt{n}\tilde{S}(v)$. Assuming the data are normally distributed, the asymptotic variance of $\tilde{S}(v)$ is

$$\text{Var}\{\tilde{S}(v)\} = \text{Var}\{\sqrt{\tilde{S}^2(v)}\} = (S^2(v)/2+1)/n. \tag{2.6}$$

Thus, the standard deviation of $\sqrt{n}\tilde{S}(v)$ is $\sqrt{S^2(v)/2+1}$. This will be termed as the parametric method.

In the case of no normalization, we assume the standard deviation of $\tilde{S}(v)$ is 1. This does not account for the sampling heterogeneity caused by estimating the variance of the effect size statistic, but still yields a valid bootstrap estimator for the SCIs. We will call this the "none" method.

2.3 Simulation Scenarios

To evaluate the proposed bootstrap procedure, we conduct a basic simulation with 1000 bootstraps and 1000 simulations for a two-sample test at each location in a 1-dimensional image with independent voxels. We first generate $X \in \mathbb{R}^n$ according to a Bernoulli random variable with parameter p, and the true effect size $S(v)$ as an independent uniform variable outside of the simulation. We set

$$\beta(v) = S(v) \times \frac{1}{\sqrt{p(1-p)}}. \tag{2.7}$$

This value of $\beta(v)$ was selected so that the true effect size index was equal to $S(v)$, the true effect size. We then set our outcome

$$Y(v) = X\beta(v) + \varepsilon(v), \tag{2.8}$$

where $\varepsilon(v) \sim N(0,1)$. We also consider the case where $\varepsilon(v) \sim Gamma(1,1)$, to evaluate the method's ability to perform under a skewed distribution.

The sample sizes n we consider are 50, 175, 300, 400, and 500. We also consider two cases of different amounts of voxels, V = 100 and V = 500, to evaluate the method's performance for larger images. Finally, we consider both cases of robustness of the standard error (SE) and RESI estimates computed, where robust (sandwich) SEs are used, and where they are not used. All combinations of these different scenarios are shown in table 2.1.

Voxel amount	Error distribution	Robust	Method
100	Normal	T	None
500	Normal	T	None
100	Gamma	T	None
500	Gamma	T	None
100	Normal	F	None
500	Normal	F	None
100	Gamma	F	None
500	Gamma	F	None
100	Normal	T	Parametric
500	Normal	T	Parametric
100	Gamma	T	Parametric
500	Gamma	T	Parametric
100	Normal	F	Parametric
500	Normal	F	Parametric
100	Gamma	F	Parametric
500	Gamma	F	Parametric

Table 2.1: This table illustrates the 16 different simulation scenarios we evaluated, under each sample size n.

2.4 Results

From Figure 2.1, we can see that there is not a clear difference between both normalization methods. There also does not seem to be a difference between both robustness cases. Comparing the normal vs gamma error terms, it seems that when the error term is skewed, the method performs poorly at low sample sizes ($n = 50$), but when the error term is normally distributed, the target coverage rate is achieved. Comparing voxel amount, there is no clear difference for the normal error term, but it causes the gamma error case to have much lower coverage at $n = 50$ when increasing to 500 voxels. Interestingly, the case of V = 500, $n = 50$, no robustness, and gamma error term, the coverage is lowest, at 0.892. Overall, as n approaches 500, all cases appear to converge to the target coverage rate. We do note that the coverage does appear to be high, seeming to increase above the target rate with sample size.

From Figure 2.2, we can see that across all cases, the parametric method creates SCIs with shorter widths compared to the "none" method. This may be because the parametric method accounts for the sampling heterogeneity that is caused by estimating the variance of the effect size statistic. The use of robust SEs has no effect on the width. The use of a gamma distributed error creates wider SCIs compared to the normal

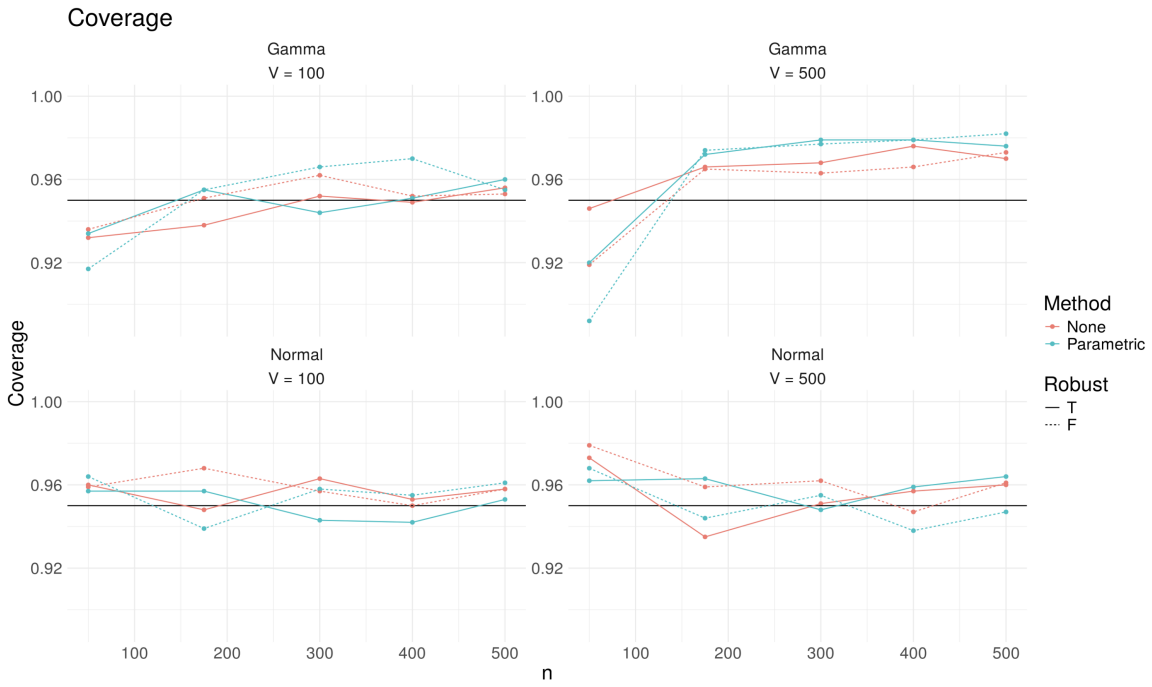

Figure 2.1: Plots of the coverage of the SCIs for values of n ranging from 50 to 500, for both cases of the error term, gamma or normally distributed, and both cases of the amount of voxels, 100 and 500. For each plot, both normalization methods are shown; the parametric method is shown in blue and the ”none” method is shown in red. The case where robust SEs are used is shown with the solid line, and the dashed lines indicate non-robust SEs. The solid black line indicates the target coverage rate, 0.95.

case. Additionally, increasing the amount of voxels causes the SCIs to be wider. These trends make sense as the width of the SCIs need to increase to accommodate for the skewedness and to capture the true effect size for more voxels. The differences we see are more prominent at lower sample sizes, but they diminish as the sample size increases. This decrease in difference appears to occur slower for the case of V = 100 and normal error term, and slowest for the case of V = 500 and gamma error term.

Although the coverage rates for the ”none” and parametric methods were similar, the mean width of the ”none” method was higher than the parametric method. Combining the results from both graphs, we can conclude that the proposed method performs well across different image sizes, but the performance slightly decreases with increase in skewedness. The robustness may not be very influential here.

Based on the results, the best choice of method would be the parametric method with the robust SEs, because although the coverage is similar with the ”none” method, the width of the SCIs is narrower using the parametric normalization. Even though the robust SEs did not appear to affect the width or coverage, it would still be safest to use due to its robustness. The weakest choice would consequently be the ”none” method with non-robust SEs, as this method yielded the wider SCIs. Additionally, the use of non-robust SEs would make the SCIs susceptible to heteroscedasticity.

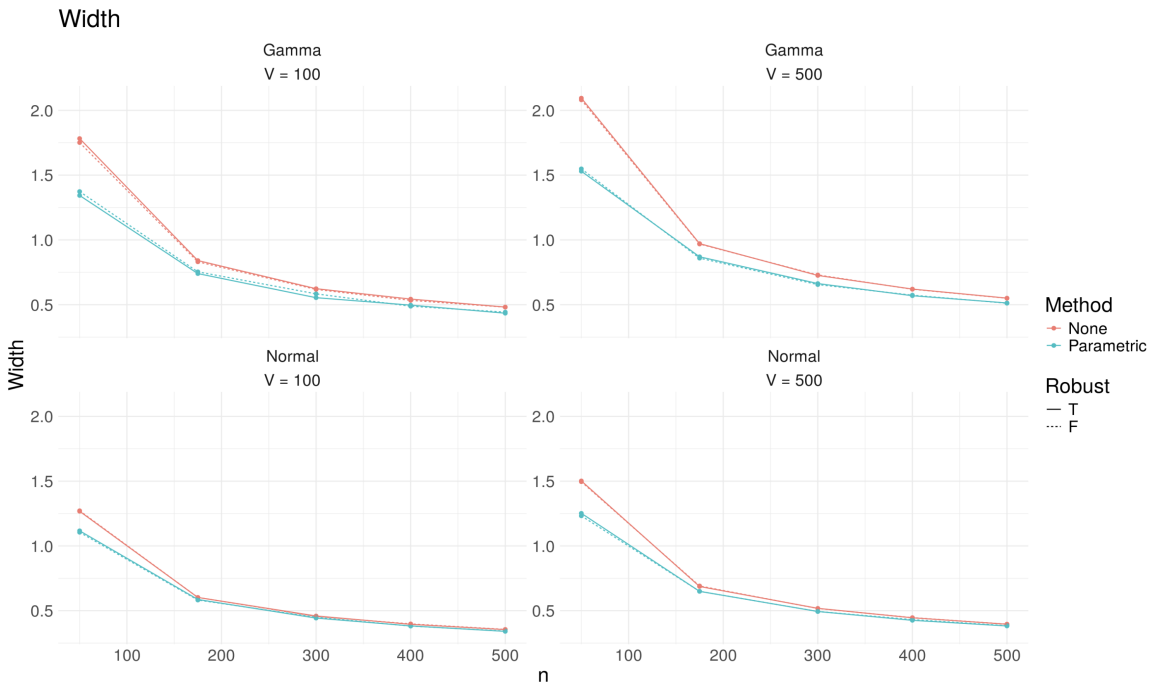

Figure 2.2: Plots of the mean width of the SCIs for values of n ranging from 50 to 500, for both cases of the error term, gamma or normally distributed, and both cases of the amount of voxels, 100 and 500. For each plot, both normalization methods are shown; the parametric method is shown in blue and the ”none” method is shown in red. The case where robust SEs are used is shown with the solid line, and the dashed lines indicate non-robust SEs.

CHAPTER 3

Discussion

3.1 Application

The simulation scenarios show that the proposed method creates SCIs that achieve the target coverage rate under many different conditions such as parametric/no SE normalization, normally/gamma distributed error terms, use of robust and non-robust SEs, and different voxel amounts. This study is the first step in creating an established method of using confidence sets for effect size-based inference. This method is generally defined for different types of statistics, but can still be extended to functional data and areas that focus on multivariate outcomes, such as genomics and imaging. Based on this general approach, we can also use this for noncontinuous terms like factor variables and nonlinear terms such as cubic splines in multiple regression.

3.2 Future Directions

Our next steps involve integrating this method into the RESI R package and developing a visualization function to display SCIs. This will aid users in interpreting regions of the image that exceed or fall short of the effect size threshold at a specified confidence level. Additionally, we plan to evaluate this new procedure under 2-dimensional and 3-dimensional images, to assess its performance in real-world applications.

Furthermore, we aim to extend this procedure to longitudinal data analysis by running additional simulation scenarios that include different working covariances. Thus, the covariance of $Y(v)$ will need to include spatial as well as temporal covariance. We will evaluate the efficiency of the different covariances while also maintaining key metrics such as SCI coverage and width. In conclusion, this general approach addresses the need of an alternative to hypothesis testing for neuroimaging inference, but can also be applied to areas outside of neuroimaging.

References

Bowring, A., Telschow, F., Schwartzman, A., and Nichols, T. E. (2019). Spatial confidence sets for raw effect size images. *NeuroImage*, page 116187.

Bowring, A., Telschow, F. J. E., Schwartzman, A., and Nichols, T. E. (2021). Confidence Sets for Cohen's d effect size images. *NeuroImage*, 226:117477.

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*.

Friston, K. J., Holmes, A., Poline, J.-B., Price, C. J., and Frith, C. D. (1996). Detecting activations in PET and fMRI: levels of inference and power. *Neuroimage*, 4(3):223–235. Number: 3.

Friston, K. J., Worsley, K. J., Frackowiak, R. S., Mazziotta, J. C., and Evans, A. C. (1994). Assessing the significance of focal activations using their spatial extent. *Human brain mapping*, 1(3):210–220. Number: 3.

Gao, X., Konietschke, F., and Li, Q. (2021). On the admissibility of simultaneous bootstrap confidence intervals. *Symmetry*, 13(7):1212.

Hall, P. (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.

Jones, M., Kang, K., and Vandekar, S. (2023). RESI: An R Package for Robust Effect Sizes. arXiv:2302.12345 [stat].

Kang, K., Jones, M. T., Armstrong, K., Avery, S., McHugo, M., Heckers, S., and Vandekar, S. (2023). Accurate Confidence and Bayesian Interval Estimation for Non-centrality Parameters and Effect Size Indices. *Psychometrika*.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., Moore, L. A., Conan, G. M., Uriarte, J., Snider, K., Lynch, B. J., Wilgenbusch, J. C., Pengo, T., Tam, A., Chen, J., Newbold, D. J., Zheng, A., Seider, N. A., Van, A. N., Metoki, A., Chauvin, R. J., Laumann, T. O., Greene, D. J., Petersen, S. E., Garavan, H., Thompson, W. K., Nichols, T. E., Yeo, B. T. T., Barch, D. M., Luna, B., Fair, D. A., and Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902):654–660. Number: 7902 Publisher: Nature Publishing Group.

Qi, S., Morris, R., Turner, J. A., Fu, Z., Jiang, R., Deramus, T. P., Zhi, D., Calhoun, V. D., and Sui, J. (2020). Common and unique multimodal covarying patterns in autism spectrum disorder subtypes. *Molecular autism*, 11:1–15.

Ren, J., Telschow, F. J. E., and Schwartzman, A. (2023). Inverse set estimation and inversion of simultaneous confidence intervals. arXiv:2210.03933 [math, stat].

Vandekar, S., Tao, R., and Blume, J. (2020). A Robust Effect Size Index. *Psychometrika*, 85(1):232–246.

Vandekar, S. N., Satterthwaite, T. D., Rosen, A., Ciric, R., Roalf, D. R., Ruparel, K., Gur, R. C., Gur, R. E., and Shinohara, R. T. (2018). Faster family-wise error control for neuroimaging with a parametric bootstrap. *Biostatistics*, 19(4):497–513. Number: 4.

Vandekar, S. N. and Stephens, J. (2021). Improving the replicability of neuroimaging findings by thresholding effect sizes instead of p-values. *Human Brain Mapping*, 42(8):2393–2398.

Wasserstein, R. L. and Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133. Number: 2.

Woo, C.-W., Krishnan, A., and Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: pitfalls and recommendations. *Neuroimage*, 91:412–419.